Multi-Garment Net: Learning to Dress 3D People from Images

Bharat Lal Bhatnagar Garvita Tiwari Christian Theobalt Gerard Pons-Moll
Max Planck Institute for Informatics, Saarland Informatics Campus, Germany
{bbhatnag,gtiwari,theobalt,gpons}@mpi-inf.mpg.de

Abstract

We present Multi-Garment Network (MGN), a method to predict body shape and clothing, layered on top of the SMPL [40] model from a few frames (1-8) of a video. Several experiments demonstrate that this representation allows higher level of control when compared to single mesh or voxel representations of shape. Our model allows to predict garment geometry, relate it to the body shape, and transfer it to new body shapes and poses. To train MGN, we leverage a digital wardrobe containing 712 digital garments in correspondence, obtained with a novel method to register a set of clothing templates to a dataset of real 3D scans of people in different clothing and poses. Garments from the digital wardrobe, or predicted by MGN, can be used to dress any body shape in arbitrary poses. We will make publicly available the digital wardrobe, the MGN model, and code to dress *SMPL with the garments at [1].*

1. Introduction

The 3D reconstruction and modelling of humans from images is a central problem in computer vision and graphics. Although a few recent methods [5, 3, 4, 25, 41, 51] attempt reconstruction of people with clothing, they lack realism and control. This limitation is in great part due to the fact that they use a single surface (mesh or voxels) to represent both clothing and body. Hence they can not capture the clothing separately from the subject in the image, let alone map it to a novel body shape.

In this paper, we introduce Multi-Garment Network (MGN), the first model capable of inferring human body and layered garments on top as separate meshes from images directly. As illustrated in Fig. 1 this new representation allows full control over body shape, texture and geometry of clothing and opens the door to a range of applications in VR/AR, entertainment, cinematography and virtual try-on.

Compared to previous work, MGN produces reconstructions of higher visual quality, and allows for more control: 1) we can infer the 3D clothing from one subject, and dress a second subject with it, (see Fig. 1, 8) and 2) we can trivially map the garment texture captured from images to any garment geometry of the same category (see Fig.7).

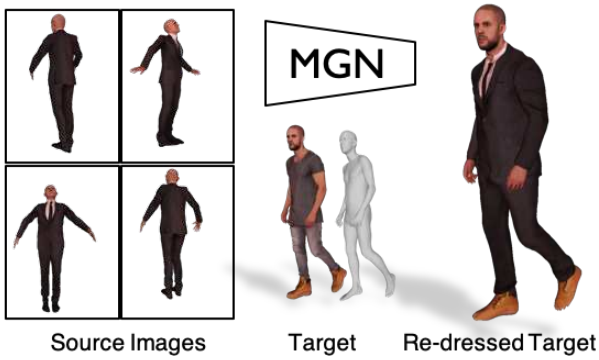

Figure 1: Garment re-targeting with Multi-Garment Network (MGN). Left to right: images from source subject, body from the target subject, target dressed with source garments. From one or more images, MGN can reconstruct the body shape and each of the garments separately. We can transfer the predicted garments to a novel body including geometry and texture.

To achieve such level of control, we address two major challenges: learning per-garment models from 3D scans of people in clothing, and learning to reconstruct them from images. We define a discrete set of garment templates (according to the categories long/short shirt, long/short pants and coat) and register, for every category, a single template to each of the scan instances, which we automatically segmented into clothing parts and skin. Since garment geometry varies significantly within one category (*e.g.* different shapes, sleeve lengths), we first minimize the distance between template and the scan boundaries, while trying to preserve the Laplacian of the template surface. This initialization step only requires solving a linear system, and nicely stretches and compresses the template globally, which we found crucial to make subsequent non-rigid registration work. Using this, we compile a *digital wardrobe* of real 3D garments worn by people, (see Fig. 3). From such registrations, we learn a vertex based PCA model per garment. Since garments are naturally associated with the underlying SMPL body model, we can transfer them to different body shapes, and re-pose them using SMPL. From

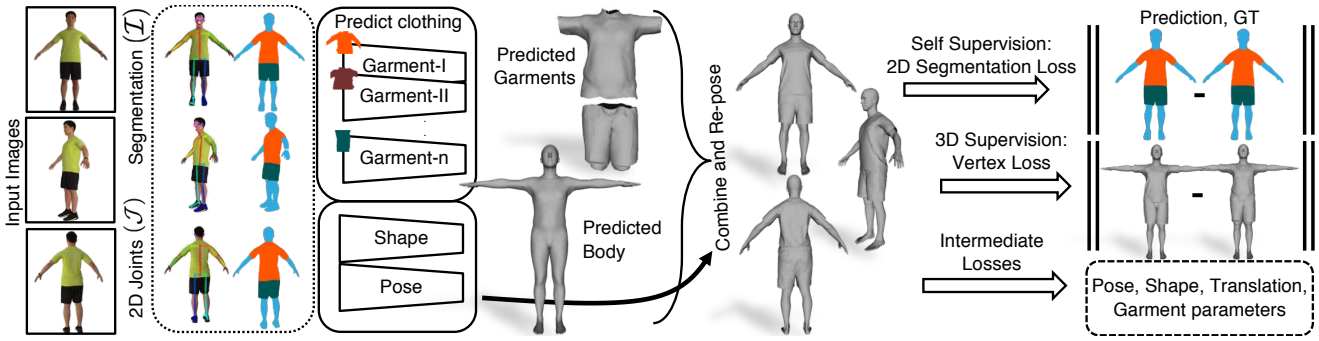

Figure 2: Overview of our approach. Given a small number of RGB frames (currently 8), we pre-compute semantically segmented images ($\mathcal{I}$) and $2D$ Joints ($\mathcal{J}$). Our Multi-Garment Network (MGN), takes $\{\mathcal{I}, \mathcal{J}\}$ as input and infers separable garments and the underlying human shape in a canonical pose. We repose these predictions using our per-frame pose predictions. We train MGN with a combination of 2D and 3D supervision. The 2D supervision can be used for online refinement at test time.

the digital wardrobe, MGN is trained to predict, given one or more images of the person, the body pose and shape parameters, the PCA coefficients of each of the garments, and a displacement field on top of PCA that encodes clothing detail. At test time, we refine this bottom-up estimates with a new top-down objective that forces projected garments and skin to explain the input semantic segmentation. This allows more fine-grained image matching as compared to standard silhouette matching. Our contributions can be summarized as:

- A novel data driven method to infer, for the first time, separate body shape and clothing from just images (few RGB images of a person rotating in front of the camera).
- A robust pipeline for 3D scan segmentation and registration of garments. To the best of our knowledge, there are no existing works capable of automatically registering a single garment template set to multiple scans of real people with clothing.
- A novel top-down objective function that forces the predicted garments and body to fit the input semantic segmentation images.
- We demonstrate several applications that were not previously possible such as dressing avatars with predicted $3D$ garments from images, and transfer of garment texture and geometry.
- We will make publicly available the MGN to predict $3D$ clothing from images, the digital wardrobe, as well as code to "dress" SMPL with it.

2. Related Work

In this section we discuss the two branches of work most related to our method, namely *capture* of clothing and body shape and *data-driven* clothing models.

Performance Capture. The classical approach to bring dynamic sequences into correspondence is to deform meshes non-rigidly [11, 18, 10] or volumetric shape representations [28, 2] to fit multiple image silhouettes. Without a pre-scanned template, fusion [30, 29, 55, 43, 58] trackers incrementally fuse geometry and appearance [66] to build the template on the fly. Although flexible, these require multi-view [57, 37, 14], one or more depth cameras [19, 45], or require the subject to stand still while turning the cameras around them [53, 38, 63, 16]. From RGB video, Habermann *et al.*[25] introduced a real time tracking system to capture non-rigid clothing dynamics. Very recently, SimulCap [59] allows multi-part tracking of human performances from a depth camera.

Body and cloth capture from images and depth. Since current statistical models can not represent clothing, most works [7, 26, 40, 68, 48, 32, 22, 67, 31, 50, 8, 33, 44, 46] are restricted to inferring body shape alone. Model fits have been used to virtually dress and manipulate people's shape and clothing in images [50, 67, 62, 36]. None of these approaches recover 3D clothing. Estimating body shape and clothing from an image has been attempted in [24, 12], but it does not separate clothing from body and requires manual intervention [65, 49] . Given a depth camera, Chen *et al.* [13] retrieve similar looking synthetic clothing templates from a database. Daněřek *et al.* [9] use physics based simulation to train a CNN but do not estimate garment and body jointly, require pre-specified garment type, and the results can only be as good as the synthetic data.

Closer to ours is the work of Alldieck *et al.* [3, 5, 6] which reconstructs, from a single image or a video, clothing and hair as displacements on top of SMPL, but can not separate garments from body, and can not transfer clothing to new subjects. In stark contrast to [3], we register the scan garments (matching boundaries) and body separately, which allows us to learn the mapping from images to a multi-layer representation of people.

Data-driven clothing. A common strategy to learn ef-

ficient data-driven models is to use off-line simulations [17, 34, 21, 54, 52, 23] for generating data. These approaches often lack realism when compared to models trained using real data. Very few approaches have shown models learned from real data. Given a dynamic scan sequence, Neophytou *et al.*[42] learn a two layer model (body and clothing) and use it to dress novel shapes. A similar model has been recently proposed [61], where the clothing layer is associated to the body in a fuzzy fashion. Other methods [60, 64] focus explicitly on estimating the body shape under clothing. Like these methods, we treat the underlying body shape as a layer, but unlike them, we segment out the different garments allowing sharp boundaries and more control. For garment registration, we build on the ideas of ClothCap [47], which can register a *subject specific* multi-part model to a 4D scan sequence. By contrast, we register a single template set to multiple scan instances– varying in garment geometry, subject identity and pose, which requires a new solution. Most importantly, unlike all previous work [35, 47, 61], we learn per-garment models and train a CNN to predict body shape and garment geometry directly from images.

3. Method

In order to learn a model to predict body shape and garment geometry directly from images, we process a dataset of 356 scans of people in varied clothing, poses and shapes. Our data pre-processing (Sec. 3.1) consists of the following steps: SMPL registration to the scans, body aware scan segmentation and template registration. We obtain, for every scan, the underlying body shape, and the garments of the person registered to one of the 5 garment template categories: shirt, t-shirt, coat, short-pants, long-pants. The obtained digital wardrobe is illustrated in Fig. 3. The garment templates are defined as regions on the SMPL surface; the original shape follows a human body, but it deforms to fit each of the scan instances after registration. Since garment registrations are naturally associated to the body represented with SMPL, they can be easily reposed to arbitrary poses. With this data, we train our Multi-Garment Network to estimate the body shape and garments from one or more images of a person, see Sec. 3.2.

3.1. Data Pre-Processing: Scan Segmentation and Registration

Unlike ClothCap [47] which registers a template to a $4D$ scan sequence of a *single subject*, our task is to register single template across instances of varying styles, geometries, body shapes and poses. Since our registration follows the ideas of [47], we describe the main differences here.

Body-Aware Scan Segmentation We first automatically segment the scans into three regions: skin, upper-clothes and pants (we annotate the garments present for every scan). Since even SOTA image semantic segmentation [20] is inaccurate, naive lifting to 3D is not sufficient. Hence, we incorporate *body specific garment priors* and segment scans by solving an MRF on the UV-map of the SMPL surface after non-rigid alignment.

A garment prior (for garment g) derives from a set of labels $\mathbf{l}_g^i \in \{0, 1\}$ indicating the vertices $\mathbf{v}_i \in \mathcal{S}$ of SMPL that are likely to overlap with the garment. The aim is to penalize labelling vertices as g outside this region, see Fig 4. Since garment geometry varies significantly within one category (*e.g.* t-shirts of different sleeve lengths), we define a cost increasing with the geodesic distance $\mathrm{dist}_{\mathrm{geo}}(\mathbf{v}) : \mathcal{S} \mapsto \mathbb{R}$ from the garment region boundary – efficiently computed based on heat flow [15]. Conversely, we define a similar penalty for labeling vertices in the garment region with a label different than g. As data terms, we incorporate CNN based semantic segmentation [20], and appearance terms based Gaussian Mixture Models in *La* color space. The influence of each term is illustrated in Fig. 4, for more details we refer to the supp. mat.

After solving the MRF on the SMPL UV map, we can segment the scans into 3 parts by transferring the labels from the SMPL registration to the scan.

Garment Template We build our garment template on top of SMPL+D, $M(\cdot)$, which represents the human body as a parametric function of pose($\boldsymbol{\theta}$), shape($\boldsymbol{\beta}$), global translation($\boldsymbol{t}$) and optional per-vertex displacements ($\mathbf{D}$):

$$M(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D}) = W(T(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D}), J(\boldsymbol{\beta}), \boldsymbol{\theta}, \mathbf{W}) \quad (1)$$

$$T(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D}) = \mathbf{T} + B_s(\boldsymbol{\beta}) + B_p(\boldsymbol{\theta}) + \mathbf{D}. \quad (2)$$

The basic principle of SMPL is to apply a series of linear displacements to a base mesh $\mathbf{T}$ with n vertices in a T-pose, and then apply standard skinning $W(\cdot)$. Specifically, $B_p(\cdot)$ models pose-dependent deformations of a skeleton J, and $B_s(\cdot)$ models the shape dependent deformations. $\mathbf{W}$ represents the blend weights.

For each garment class g we define a template mesh, $\mathbf{G}^g$ in T-pose, which we subsequently register to explain the scan garments. We define $\mathbf{I}^g \in \mathbb{Z}^{m_g \times n}$ as an indicator matrix, with $\mathbf{I}^g_{i,j} = 1$ if garment g vertex $i \in \{1 \dots m_g\}$ is associated with body shape vertex $j \in \{1 \dots n\}$. In our experiments, we associate a single body shape vertex to each garment vertex. We compute displacements to the corresponding SMPL body shape $\boldsymbol{\beta}^g$ under the garment as

$$\mathbf{D}^g = \mathbf{G}^g - \mathbf{I}^g T(\boldsymbol{\beta}^g, \mathbf{0}_\theta, \mathbf{0}_{\mathbf{D}}) \quad (3)$$

Consequently, we can obtain the garment shape (unposed), T^g for a new shape $\boldsymbol{\beta}$ and pose $\boldsymbol{\theta}$ as

$$T^g(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D}^g) = \mathbf{I}^g T(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{0}) + \mathbf{D}^g \quad (4)$$

To pose the vertices of a garment, each vertex uses the skinning function in Eq. 1 of the associated SMPL body vertex.

$$G(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D}^g) = W(T^g(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D}^g), J(\boldsymbol{\beta}), \boldsymbol{\theta}, \mathbf{W}) \quad (5)$$

Figure 3: Digital 3D wardrobe. We use our proposed multi-mesh registration approach to register garments present in the scans (left) to fixed garment templates. This allows us to build a digital wardrobe and dress arbitrary subjects (center) by picking the garments (marked) from the wardrobe.

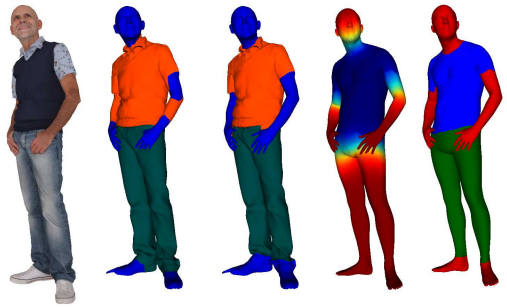

Figure 4: Left to right: Scan, segmentation with MRF and CNN unaries, MRF with CNN unaries + garment prior + appearance terms, the garment(t-shirt) prior based on geodesics and the template. Notice how the garment prior is crucial to obtain robust results.

Garment Registration Given the segmented scans, we non-rigidly register the body and garment templates (upper-clothes, lower-clothes) to scans using the multi-part alignment proposed in [47]. The challenging part is that garment geometries vary significantly across instances, which makes the multi-part registration fail (see supplementary). Hence, we first initialize by deforming the vertices of each garment template with the shape and pose of SMPL registrations, obtaining deformed vertices $\mathbf{G}^g_{\text{init}}$. Note that since the vertices defining each garment template are fixed, the clothing boundaries of the initially deformed garment template will not match the scan boundaries. In order to globally deform the template to match the clothing boundaries in a single shot, we define an objective function based on Laplacian deformation [56].

Let $\mathbf{L}^g \in \mathbb{R}^{m_g \times m_g}$ be the graph Laplacian of the garment mesh, and $\mathbf{\Delta}_{\text{init}} \in \mathbb{R}^{m_g \times 3}$ the differential coordinates of the initially deformed garment template $\mathbf{\Delta}_{\text{init}} = \mathbf{L}\,\mathbf{G}^g_{\text{init}}$. For every vertex $\mathbf{s}_i \in \mathcal{S}_b$ in a scan boundary $\mathcal{S}_b$, we find its closest vertex in the corresponding template garment boundary, obtaining a matrix of scan points $\mathbf{q}_{1:C} = \{\mathbf{q}_1, \ldots, \mathbf{q}_C\}$ with corresponding template vertex indices $j_{1:C}$. Let $\mathbf{I}_{C \times m_g}$ be a selector matrix indicating the indices in the template corresponding to each $\mathbf{q}_i$. With this, we minimize the following least squares problem:

$$\begin{bmatrix} \mathbf{L}^g \\ w\mathbf{I}_{C \times m_g} \end{bmatrix} \mathbf{G}^g = \begin{bmatrix} \mathbf{\Delta}_{\text{init}} \\ w\mathbf{q}_{1:C} \end{bmatrix} \tag{6}$$

with respect to the template garment vertices $\mathbf{G}^g$, where the first block $\mathbf{L}^g\mathbf{G}^g = \mathbf{\Delta}_{\text{init}}$ forces the solution to keep the local surface structure, while the second block $w\mathbf{I}_{C \times m_g}\mathbf{G}^g = w\mathbf{q}_{1:C}$ makes the boundaries match. The nice property of the linear system solve is that the garment template globally stretches or compresses to match the scan garment boundaries, which would take many iterations of non-linear non-rigid registration [47] with the risk of converging to bad local minima. After this initialization, we non-linearly register each garment $\mathbf{G}^g$ to fit the scan surface. We build on top of the proposed multi-part registration in [47] and propose additional loss terms on garment vertices, $v_k \in \mathbf{G}^g$, to facilitate better garment unposing, E_{unpose}, and minimize interpenetration, E_{interp}, with the underlying SMPL body surface, $\mathcal{S}$.

$$E_{\text{interp}} = \sum_g \sum_{\mathbf{v}_k \in \mathbf{G}^g} d(\mathbf{v}_k, \mathcal{S}) \tag{7}$$

$$d(\mathbf{x}, \mathcal{S}) = \begin{cases} 0, & \text{if } \mathbf{x} \text{ outside } \mathcal{S} \\ w * |\mathbf{x} - \mathbf{y}|_2, & \text{if } \mathbf{x} \text{ inside } \mathcal{S} \end{cases} \tag{8}$$

where w is a constant ($w = 25$ in our experiments), $\mathbf{v}_k$ is the k^{th} vertex of $\mathbf{G}^g$ and $\mathbf{y}$ is the point closest to $\mathbf{x}$ on $\mathcal{S}$.

Our garment formulation allows us to freely repose the garment vertices. We can use this to our advantage for applications such as animating clothed virtual avatars, garment re-targeting etc. However, posing is highly non-linear and can lead to undesired artefacts, specially when re-targeting garments across subjects with very different poses.

Since we re-target the garments in unposed space, we reduce distortion by forcing distances from garment vertices to the body to be preserved after unposing:

$$E_{\text{unpose}} = \sum_g \sum_{\mathbf{v}_k \in \mathbf{G}^g} (d(\mathbf{v}_k, \mathcal{S}) - d(\mathbf{v}_k^0, \mathcal{S}^0))^2 \quad (9)$$

where $d(\mathbf{x}, \mathcal{S})$ is the L_2 distance between point $\mathbf{x}$ and surface $\mathcal{S}$. $\mathbf{v}_k^0$ and $\mathcal{S}^0$ denote garment vertex and body surface in unposed space, using Eq. 5 and 1 respectively.

Dressing SMPL The SMPL model has proven very useful for modelling unclothed shapes. Our idea is to build a wardrobe of digital clothing compatible with SMPL to model clothed subjects. To this end we propose a simple extension that allows to *dress* SMPL. Given a garment $\mathbf{G}^g$, we use Eq. 3, 4, 5 to pose and skin the garment vertices. The dressed body including body shape (encoded as G_1) will be given by stacking the L individual garment vertices $[G_1(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D}_1)^T, \ldots, G_L(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D}_L)^T]^T$. We define the function $C(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{D})$ which returns the posed and shaped vertices for the skin, and each of the garments combined. See Fig. 5 and supplementary for results on re-targeting garments using MGN across different SMPL bodies.

3.2. From Images to Garments

From registrations, we learn a shape space of garments, and generate a synthetic training dataset with pairs of images and body+3D garment pairs. From this data we train MGN:*Multi-Garment Net*, which maps images to 3D garments and body shape.

Garment Shape Space In order to factor out pose deformations from garment shape, we "unpose" the j^{th} garment registrations $\mathbf{G}_j^g \in \mathbb{R}^{m_g \times 3}$, similar to [64, 47]. Since the garments of each category are all in correspondence, we can easily compute PCA directly on the unposed vertices to obtain pose-invariant shape basis ($\mathbf{B}^g$). Using this, we encode a garment shape using 35 components $\boldsymbol{z}^g \in \mathbb{R}^{35}$, plus a residual vector of offsets $\mathbf{D}_j^{\text{hf},g}$, mathematically: $\mathbf{G}_j^g = \mathbf{B}^g \mathbf{z}_j^g + \mathbf{D}_j^{\text{hf},g}$. From each scan, we also extract the body shape *under* clothing similarly as in [64], which is essential to re-target a garment from one body to another.

MGN: Multi-Garment Net The input to the model is a set of semantically segmented images, $\mathcal{I} = \{\mathbf{I}_0, \mathbf{I}_1, ..., \mathbf{I}_F - 1\}$, and corresponding $2D$ joint estimates, $\mathcal{J} = \{\mathbf{J}_0, \mathbf{J}_1, ..., \mathbf{J}_F - 1\}$, where F is the number of images used to make the prediction. Following [20, 3], we abstract away the appearance information in RGB images and extract semantic garment segmentation [20] to reduce the risk of over-fitting, albeit at the cost of disregarding useful shading signal. For simplicity, let now $\boldsymbol{\theta}$ denote both the joint angles $\boldsymbol{\theta}$ and translation $\boldsymbol{t}$.

The base network, f_w, maps the 2D poses $\mathcal{J}$, and image segmentations $\mathcal{I}$, to per frame latent code ($\boldsymbol{l}_\mathcal{P}$) corresponding to 3D poses

$$\boldsymbol{l}_\mathcal{P} = f_w^\theta(\mathcal{I}, \mathcal{J}), \quad (10)$$

and to a common latent code corresponding to body shape ($\boldsymbol{l}_\beta$) and garments ($\boldsymbol{l}_\mathcal{G}$) by averaging the per frame codes

$$\boldsymbol{l}_\boldsymbol{\beta}, \boldsymbol{l}_\mathcal{G} = \frac{1}{F} \sum_{f=0}^{F-1} f_w^{\boldsymbol{\beta},\mathcal{G}}(\mathbf{I}_f, \mathbf{J}_f). \quad (11)$$

For each garment class, we train separate branches, $M_w^g(\cdot)$, to map the latent code $\boldsymbol{l}_\mathcal{G}$ to the un-posed garment $\mathbf{G}^g$, which itself is reconstructed from low-frequency PCA coefficients $\mathbf{z}^g$, plus $\mathbf{D}^{\text{hf,g}}$ encoding high-frequency displacements

$$M_w^g(\boldsymbol{l}_\mathcal{G}, \mathbf{B}^g) = \mathbf{G}^g = \mathbf{B}^g \boldsymbol{z}^g + \mathbf{D}^{\text{hf},g}. \quad (12)$$

From the shape and pose latent codes $\boldsymbol{l}_\boldsymbol{\beta}, \boldsymbol{l}_\boldsymbol{\theta}$, we predict body shape parameters $\boldsymbol{\beta}$ and pose $\boldsymbol{\theta}$ respectively, using a fully connected layer. Using the predicted body shape $\boldsymbol{\beta}$ and geometry $M_w^g(\boldsymbol{l}_\mathcal{G}, \mathbf{B}^g)$ we compute displacements as in Eq. 3:

$$\mathbf{D}^g = M_w^g(\boldsymbol{l}_\mathcal{G}, \mathbf{B}^g) - \mathbf{I}^g T(\boldsymbol{\beta}, \mathbf{0}_\theta, \mathbf{0}_\mathbf{D}). \quad (13)$$

Consequently, the final predicted $3D$ vertices posed for the f^{th} frame are obtained with $C(\boldsymbol{\beta}, \boldsymbol{\theta}_f, \mathbf{D})$, from which we render 2D segmentation masks

$$\mathbf{R}_f = R(C(\boldsymbol{\beta}, \boldsymbol{\theta}_f, \mathbf{D}), c), \quad (14)$$

where $R(\cdot)$ is a differentiable renderer [27], $\mathbf{R}_f$ the rendered semantic segmentation image for frame f, and c denotes the camera parameters that are assumed fixed while the person moves. The rendering layer in Eq. (14) allows us to compare predictions against the input images. Since MGN predicts body and garments separately, we can predict a semantic segmentation image, leading to a more fine-grained $2D$ loss, which is not possible using a single mesh surface representation [3]. Note that Eq. 14 allows to train with self-supervision.

3.3. Loss functions

The proposed approach can be trained with $3D$ supervision on vertex coordinates, and with self supervision in the form of $2D$ segmented images. We use upper-hat for variables that are known and used for supervision during training. We use the following losses to train the network in an end to end fashion:

- $3D$ vertex loss in the canonical T-pose ($\boldsymbol{\theta} = \mathbf{0}_\boldsymbol{\theta}$):

$$\mathcal{L}_{\mathbf{0}_\boldsymbol{\theta}}^{3D} = ||C(\boldsymbol{\beta}, \mathbf{0}_\boldsymbol{\theta}, \mathbf{D}) - C(\hat{\boldsymbol{\beta}}, \mathbf{0}_\boldsymbol{\theta}, \hat{\mathbf{D}})||^2, \quad (15)$$

  where, $\mathbf{0}_\boldsymbol{\theta}$ represents zero-vector corresponding to zero pose.

Figure 5: Dressing SMPL with just images. We use MGN to extract garments from the images of a source subject (middle) and use the inferred 3D garments to dress arbitrary human bodies in various poses from SMPL shape subjects. The two sets correspond to male (left) and female (right) body shapes respectively.

- $3D$ vertex loss in posed space:

$$\mathcal{L}_{\mathcal{P}}^{3D} = \sum_{f=0}^{F-1} ||C(\boldsymbol{\beta}, \boldsymbol{\theta}_f, \mathbf{D}) - C(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}_f, \hat{\mathbf{D}})||^2 \quad (16)$$

- $2D$ segmentation loss: Unlike [3] we do not optimize silhouette overlap, instead we jointly optimize the projected per-garment segmentation against the input segmentation mask. This ensures that each garment explains its corresponding mask in the image:

$$\mathcal{L}_{seg}^{2D} = \sum_{f=0}^{F-1} ||\mathbf{R}_f - \mathbf{I}_f||^2, \quad (17)$$

- Intermediate losses: We further impose losses on intermediate pose, shape and garment parameter predictions: $\mathcal{L}_{\boldsymbol{\theta}} = \sum_{f=0}^{F-1} ||\hat{\boldsymbol{\theta}}_f - \boldsymbol{\theta}_f||^2, \mathcal{L}_{\boldsymbol{\beta}} = ||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||^2, \mathcal{L}_{\boldsymbol{z}} = \sum_{g=0}^{L-1} ||\hat{\boldsymbol{z}}^g - \boldsymbol{z}^g||^2$ where F, L are the number of images and garments respectively. $\hat{\boldsymbol{z}}$ are the ground truth PCA garment parameters. While such losses are a bit redundant, they stabilize learning.

3.4. Implementation details

Base Network (f_w^*): We use a CNN to map the input set $\{\mathcal{I}, \mathcal{J}\}$ to the body shape, pose and garment latent spaces. It consists of five, $2D$ convolutions followed by max-pooling layers. Translation invariance, unfortunately, renders CNNs unable to capture the location information of the features. In order to reproduce garment details in 3D, it is important to leverage 2D features as well as their location in the 2D image. To this end, we adopt a strategy similar to [39], where we append the pixel coordinates to the output of every CNN layer. We split the last convolutional feature maps into three parts to individuate the body shape, pose and garment information. The three branches are flattened out and we append 2D joint estimates to the pose branch. Three fully connected layers and average pooling on garment and shape latent codes, generate $\boldsymbol{l}_{\boldsymbol{\beta}}, \boldsymbol{l}_{\boldsymbol{\theta}}$ and $\boldsymbol{l}_{\mathcal{G}}$ respectively. See supplementary for more details.

Garment Network (M_w^g): We train separate garment networks for each of the garment classes. The garment network consists of two branches. The first predicts the overall mesh shape, and second one adds high frequency details. From the garment latent code ($\boldsymbol{l}_{\mathcal{G}}$), the first branch, consisting of two fully connected layers (sizes=1024, 128), regresses the PCA coefficients. Dot product of these coefficients with the PCA basis generates the base garment mesh. We use the second fully connected branch (size = m^g) to regress displacements on top of the mesh predicted in the first branch. We restrict these displacements to $\leq 1cm$ to ensure that overall shape is explained by the PCA mesh and not these displacements.

4. Dataset and Experiments

Dataset We use 356 3D scans of people with various body shapes, poses and in diverse clothing. We held out 70 scans for testing and use the rest for training. Similar to [3, 5], we also restrict our setting to the scenario where the person is turning around in front of the camera. We register the scans using multi-mesh registration, SMPL+G. This enables further data augmentation since the registered scans can now be re-posed and re-shaped.

We adopt the data pre-processing steps from [3] including the rendering and segmentation. We also acknowledge the scale ambiguity primarily present between the object size and the distance to the camera. Hence we assume that the subjects in 3D have a fixed height and regress their distance from the camera. Same as [3], we also ignore the effect of camera intrinsics.

4.1. Experiments

In this section we discuss the merits of our approach both qualitatively and quantitatively. We also show real world applications in the form of texture transfer (Fig. 7), where

Figure 6: Qualitative comparison with Alldieck *et al.*[3]. In each set we visualize 3D predictions from [3](left) and our method (right) for five test subjects. Since our approach explicitly models garment geometry, it preserves more garment details, as is evident from minimal distortions across all the subjects. For more results see supplementary.

Figure 7: Texture transfer. We model each garment class as a mesh with fixed topology and surface parameterization. This enables us to transfer texture from any garment to any other registered instance of the same class. The first column shows the source garment mesh, while the subsequent images show original and transferred garment texture registrations.

we maintain the original geometry of the source garment but map novel texture. We also show garment re-targeting from images using MGN in Fig. 8.

Qualitative comparisons: We compare our method against [3] on our scan dataset. For fair comparison we re-train the models proposed by Alldieck *et al.*[3] on our dataset and compare against our approach (Dataset used by [3] is not publicly available). Figure 6 indicates the advantage of incorporating the garment model in structured prediction over simply modelling free form displacements. Explicit garment modelling allows us to predict sharper garment boundaries and minimize distortions (see Fig. 6). More examples are shown in the supplementary material.

Quantitative Comparison: In this experiment we do a quantitative analysis of our approach against the state of the art $3D$ prediction method, [3]. We compute a symmetric error between the predicted and GT garment surfaces similar to [3]. We report per-garment error, E^g (supplementary), and overall error, i.e. mean of E^g over all the garments

$$E^g = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{|\hat{\mathbf{S}}_i^g|}\sum_{\mathbf{v}_k\in\hat{\mathbf{S}}_i^g} d(\mathbf{v}_k,\mathcal{S}_i^g)+\frac{1}{|\mathbf{S}_i^g|}\sum_{\mathbf{v}_k\in\mathbf{S}_i^g} d(\mathbf{v}_k,\hat{\mathcal{S}}_i^g)\right), \tag{18}$$

where N is the number of meshes with garment g. $\mathbf{S}_i^g$ and $\mathcal{S}_i^g$ denote the set of vertices and the surface of the i^{th} predicted mesh respectively, belonging to garment g. Operator $\hat{(.)}$ denotes GT values. $d(\mathbf{v}_k,\mathcal{S})$ computes the L_2 distance between the vertex $\mathbf{v}_k$ and surface $\mathcal{S}$.

This criterion is slightly different than [3] because we do not evaluate error on the skin parts. We reconstruct the 3D garments with mean vertex-to-surface error of 5.78 mm with 8 frames as input. We re-train octopus [3] on our dataset and the resulting error is 5.72mm.

We acknowledge the slightly better performance of [3] and attribute it to the fact that the single mesh based approaches do not bind vertices to semantic roles, i.e these approaches can pull vertices from any part of the mesh to explain $3D$ deformations where as our approach ensures that only semantically correct vertices explain the $3D$ shape.

It is also worth noting that MGN predicts garments as linear function (PCA coefficients) of latent code, whereas [3] deploys GraphCNN. PCA based formulation though easily tractable is inherently biased towards smooth results. Our work paves the way for further exploration into building garment models for modelling the variations in garment geometry over a fixed topology.

We report the results for using varying number of frames in the supplementary.

Figure 8: Garment re-targeting by MGN using 8 RGB images. In each of the three sets we show the source subject, target subject and re-targeted garments. Using MGN, we can re-target garments including both texture and geometry.

GT vs Predicted pose: The $3D$ vertex predictions are a function of pose and shape. In this experiment we do an ablation study to isolate the effect of errors in pose estimation on vertex predictions. This experiment is important to better understand the strengths and weaknesses of the proposed approach in shape estimation by marginalizing over the errors due to pose fitting. We study two scenarios, first where we predict the 3D pose and second, where we have access to GT pose. We report mean vertex-to-surface error of 5.78mm with GT poses and 11.90mm with our predicted poses.

4.2. Re-targeting

Our multi-mesh representation essentially decouples the underlying body and the garments. This opens up an interesting possibility to take garments from source subject and virtually dress a novel subject. Since the source and the target subjects could be in different poses, we first unpose the source body and garments along with the target body. We drop the $(.)^0$ notation for the unposed space in the following section for clarity. Below we propose and compare two garment re-targeting approaches. After re-targeting the target body and re-targeted garments are re-posed to their original poses.

Naive re-targeting: The simplest approach to re-target clothes from source to target is to extract the garment offsets, $\mathbf{D}^{s,g}$ from the source subject using Eq. 13 and dress a target subject using Eq. 5.

Body aware re-targeting: The naive approach is problematic because it relies on non-local pre-set vertex association between the garment and the body (I^g). This results in inaccurate association between the body blend shapes, $B_{p,s}$ and the garment vertices. This eventually leads to incorrect estimation of source offsets, $\mathbf{D}^{s,g}$ and in turn leads to higher inter-penetrations between the re-targeted garment and the body (see supplementary). In order to mitigate this issue, we compute the new k^{th} target garment vertex location, $\mathbf{v}_k^t$ as follows

$$\mathbf{v}_k^t = \mathbf{v}_k^s - \mathbf{S}_{I_k}^s + \mathbf{S}_{\mathrm{I}_k}^t \tag{19}$$

$$\mathrm{I}_k = \underset{\mathrm{I}\in[0,\,|\mathbf{S}^s|-1]}{\operatorname{argmin}} ||\mathbf{v}_k^s - \mathbf{S}_{\mathrm{I}}^s||_2, \tag{20}$$

where $\mathbf{v}_k^s$ is the source garment vertex, $\mathbf{S}_{\mathrm{I}_k}^s$ is the vertex (indexed by I_k) among the source body vertices, $\mathbf{S}^s$, closest to $\mathbf{v}_k^s$ and $\mathbf{S}_{\mathrm{I}_k}^t$ is the corresponding vertex among the target body vertices.

MGN allows us to predict separable body shape and garments in 3D, allowing us to do garment re-targeting (as described above) using just images. To the best of our knowledge this is the first method to do so. See Fig. 8 for results on garment re-targeting by MGN. See supplementary for more results.

5. Conclusion and Future Works

We introduce MGN, the first model capable of jointly reconstructing from few images, body shape and garment geometry as layered meshes. Experiments demonstrate that this representation has several benefits: it is closer to how clothing layers on top of the body in the real world, which allows control such as re-dressing novel shapes with the reconstructed clothing. Additionally, we introduce for the first time, a dataset of registered *real* garments from real scans obtained with a robust registration pipeline. When compared to more classical single mesh representations, it allows more control and qualitatively the results are very similar. In summary, we think that MGN provides a first step in a promising research direction. We will release the MGN model and the digital wardrobe to stimulate research in this direction. Further discussion on limitations and future works in supplementary.

Acknowledgements This work is partly funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 409792180 (Emmy Noether Programme, project: Real Virtual Humans) and Google Faculty Research Award. We thank twindom (https://web.twindom.com/) for providing scan data, Thiemo Alldieck for providing code for texture/segmentation stitching, and Verica Lazova for discussions.

References

[1] https://virtualhumans.mpi-inf.mpg.de/mgn/. 1

[2] Benjamin Allain, Jean-Sébastien Franco, and Edmond Boyer. An Efficient Volumetric Framework for Shape Tracking. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 268–276, Boston, United States, 2015. IEEE. 2

[3] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single RGB camera. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 5, 6, 7

[4] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Detailed human avatars from monocular video. In *International Conf. on 3D Vision*, sep 2018. 1

[5] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3D people models. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2018. 1, 2, 6

[6] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2shape: Detailed full human body geometry from a single image. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2019. 2

[7] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. SCAPE: shape completion and animation of people. In *ACM Transactions on Graphics*, volume 24, pages 408–416. ACM, 2005. 2

[8] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *European Conf. on Computer Vision*. Springer International Publishing, 2016. 2

[9] R c, Endri Dibra, C Öztireli, Remo Ziegler, and Markus Gross. Deepgarment: 3d garment shape estimation from a single image. In *Computer Graphics Forum*, volume 36, pages 269–280. Wiley Online Library, 2017. 2

[10] Cedric Cagniart, Edmond Boyer, and Slobodan Ilic. Probabilistic deformable surface tracking from multiple videos. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *European Conf. on Computer Vision*, volume 6314 of *Lecture Notes in Computer Science*, pages 326–339, Heraklion, Greece, 2010. Springer. 2

[11] Joel Carranza, Christian Theobalt, Marcus A Magnor, and Hans-Peter Seidel. Free-viewpoint video of human actors. In *ACM Transactions on Graphics*, volume 22, pages 569–577. ACM, 2003. 2

[12] Xiaowu Chen, Yu Guo, Bin Zhou, and Qinping Zhao. Deformable model for estimating clothed and naked human shapes from a single image. *The Visual Computer*, 29(11):1187–1196, 2013. 2

[13] Xiaowu Chen, Bin Zhou, Feixiang Lu, Lin Wang, Lang Bi, and Ping Tan. Garment modeling with a depth camera. *ACM Transactions on Graphics*, 34(6):203, 2015. 2

[14] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics*, 34(4):69, 2015. 2

[15] Keenan Crane, Clarisse Weischedel, and Max Wardetzky. Geodesics in heat: A new approach to computing distance based on heat flow. *ACM Transactions on Graphics (TOG)*, 32(5):152, 2013. 3

[16] Yan Cui, Will Chang, Tobias Nöll, and Didier Stricker. Kinectavatar: fully automatic body capture using a single kinect. In *Asian Conf. on Computer Vision*, pages 133–147, 2012. 2

[17] Edilson de Aguiar, Leonid Sigal, Adrien Treuille, and Jessica K. Hodgins. Stable spaces for real-time clothing. *ACM Trans. Graph.*, 29(4):106:1–106:9, July 2010. 3

[18] Edilson De Aguiar, Carsten Stoll, Christian Theobalt, Naveed Ahmed, Hans-Peter Seidel, and Sebastian Thrun. Performance capture from sparse multi-view video. In *ACM Transactions on Graphics*, page 98, 2008. 2

[19] Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip Davidson, Sean Ryan Fanello, Adarsh Kowdle, Sergio Orts Escolano, Christoph Rhemann, David Kim, Jonathan Taylor, et al. Fusion4d: Real-time performance capture of challenging scenes. *ACM Transactions on Graphics*, 35(4):114, 2016. 2

[20] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. Instance-level human parsing via part grouping network. In *European Conf. on Computer Vision*, 2018. 3, 5

[21] P. Guan, L. Reiss, D. Hirshberg, A. Weiss, and M. J. Black. DRAPE: DRessing Any PErson. *ACM Trans. on Graphics (Proc. SIGGRAPH)*, 31(4):35:1–35:10, July 2012. 3

[22] Peng Guan, Alexander Weiss, Alexandru O Bălan, and Michael J Black. Estimating human shape and pose from a single image. In *IEEE International Conf. on Computer Vision*, pages 1381–1388. IEEE, 2009. 2

[23] Erhan Gundogdu, Victor Constantin, Amrollah Seifoddini, Minh Dang, Mathieu Salzmann, and Pascal Fua. Garnet: A two-stream network for fast and accurate 3d cloth draping. *arXiv preprint arXiv:1811.10983*, 2018. 3

[24] Yu Guo, Xiaowu Chen, Bin Zhou, and Qinping Zhao. Clothed and naked human shapes estimation from a single image. *Computational Visual Media*, pages 43–50, 2012. 2

[25] Marc Habermann, Weipeng Xu, , Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Livecap: Real-time human performance capture from monocular video. *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, jul 2019. 1, 2

[26] Nils Hasler, Carsten Stoll, Martin Sunkel, Bodo Rosenhahn, and H-P Seidel. A statistical model of human pose and body shape. In *Computer Graphics Forum*, volume 28, pages 337–346, 2009. 2

[27] Paul Henderson and Vittorio Ferrari. Learning to generate and reconstruct 3d meshes with only 2d supervision. In *British Machine Vision Conference (BMVC)*, 2018. 5

[28] Chun-Hao Huang, Benjamin Allain, Jean-Sébastien Franco, Nassir Navab, Slobodan Ilic, and Edmond Boyer. Volumetric 3d tracking by detection. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 3862–3870, 2016. 2

[29] Matthias Innmann, Michael Zollhöfer, Matthias Nießner, Christian Theobalt, and Marc Stamminger. Volumedeform: Real-time volumetric non-rigid reconstruction. In *European Conf. on Computer Vision*, 2016. 2

[30] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, et al. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *ACM symposium on User interface software and technology*, pages 559–568. ACM, 2011. 2

[31] Arjun Jain, Thorsten Thormählen, Hans-Peter Seidel, and Christian Theobalt. Moviereshape: Tracking and reshaping of humans in videos. In *ACM Transactions on Graphics*, volume 29, page 148. ACM, 2010. 2

[32] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 8320–8329, 2018. 2

[33] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *IEEE Conf. on Computer Vision and Pattern Recognition*. IEEE Computer Society, 2018. 2

[34] Doyub Kim, Woojong Koh, Rahul Narain, Kayvon Fatahalian, Adrien Treuille, and James F. O'Brien. Near-exhaustive precomputation of secondary cloth effects. *ACM Transactions on Graphics*, 32(4):87:1–7, July 2013. Proceedings of ACM SIGGRAPH 2013, Anaheim. 3

[35] Zorah Lahner, Daniel Cremers, and Tony Tung. Deepwrinkles: Accurate and realistic clothing modeling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 667–684, 2018. 3

[36] Christoph Lassner, Gerard Pons-Moll, and Peter V. Gehler. A generative model of people in clothing. In *Proceedings IEEE International Conference on Computer Vision (ICCV)*, Piscataway, NJ, USA, oct 2017. IEEE. 2

[37] Vincent Leroy, Jean-Sébastien Franco, and Edmond Boyer. Multi-View Dynamic Shape Refinement Using Local Temporal Integration. In *IEEE International Conf. on Computer Vision*, Venice, Italy, 2017. 2

[38] Hao Li, Etienne Vouga, Anton Gudym, Linjie Luo, Jonathan T Barron, and Gleb Gusev. 3d self-portraits. *ACM Transactions on Graphics*, 32(6):187, 2013. 2

[39] Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. An intriguing failing of convolutional neural networks and the coordconv solution. In *Proceedings of the 32Nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 9628–9639, USA, 2018. Curran Associates Inc. 6

[40] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6):248:1–248:16, 2015. 1, 2

[41] Ryota Natsume, Shunsuke Saito, Zeng Huang, Weikai Chen, Chongyang Ma, Hao Li, and Shigeo Morishima. Siclope: Silhouette-based clothed people. *arXiv preprint arXiv:1901.00049*, 2018. 1

[42] Alexandros Neophytou and Adrian Hilton. A layered model of human body and garment deformation. In *International Conference on 3D Vision*, 2014. 3

[43] Richard A Newcombe, Dieter Fox, and Steven M Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 343–352, 2015. 2

[44] Mohamed Omran, Christop Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *International Conf. on 3D Vision*, 2018. 2

[45] Sergio Orts-Escolano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L Davidson, Sameh Khamis, Mingsong Dou, et al. Holoportation: Virtual 3d teleportation in real-time. In *Symposium on User Interface Software and Technology*, pages 741–754. ACM, 2016. 2

[46] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2018. 2

[47] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael Black. ClothCap: Seamless 4D clothing capture and retargeting. *ACM Transactions on Graphics*, 36(4), 2017. 3, 4, 5

[48] Gerard Pons-Moll, Javier Romero, Naureen Mahmood, and Michael J Black. Dyna: a model of dynamic human shape in motion. *ACM Transactions on Graphics*, 34:120, 2015. 2

[49] Cody Robson, Ron Maharik, Alla Sheffer, and Nathan Carr. Context-aware garment modeling from sketches. *Computers & Graphics*, 35(3):604–613, 2011. 2

[50] Lorenz Rogge, Felix Klose, Michael Stengel, Martin Eisemann, and Marcus Magnor. Garment replacement in monocular video sequences. *ACM Transactions on Graphics*, 34(1):6, 2014. 2

[51] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. *arXiv preprint arXiv:1905.05172*, 2019. 1

[52] Igor Santesteban, Miguel A. Otaduy, and Dan Casas. Learning-based animation of clothing for virtual try-on. In *Computer Graphics Forum (Proc. Eurographics)*, volume 32, pages 1–8, 2019. 3

[53] Ari Shapiro, Andrew Feng, Ruizhe Wang, Hao Li, Mark Bolas, Gerard Medioni, and Evan Suma. Rapid avatar capture and simulation using commodity depth sensors. *Computer Animation and Virtual Worlds*, 25(3-4):201–211, 2014. 2

[54] Leonid Sigal, Moshe Mahler, Spencer Diaz, Kyna McIntosh, Elizabeth Carter, Timothy Richards, and Jessica Hodgins. A perceptual control space for garment simulation. In *ACM Transactions on Graphics*, 2015. 3

[55] Miroslava Slavcheva, Maximilian Baust, Daniel Cremers, and Slobodan Ilic. Killingfusion: Non-rigid 3d reconstruction without correspondences. In *IEEE Conf. on Computer Vision and Pattern Recognition*, volume 3, page 7, 2017. 2

[56] Olga Sorkine. Laplacian mesh processing. In *Eurographics (STARs)*, pages 53–70, 2005. 4

[57] Jonathan Starck and Adrian Hilton. Surface capture for performance-based animation. *IEEE Computer Graphics and Applications*, 27(3), 2007. 2

[58] Yu Tao, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Dai Quionhai, Hao Li, G. Pons-Moll, and Yebin Liu. Doublefusion: Real-time capture of human performance with inner body shape from a depth sensor. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2018. 2

[59] Yu Tao, Zerong Zheng, Yuan Zhong, Jianhui Zhao, Dai Quionhai, Gerard Pons-Moll, and Yebin Liu. Simulcap : Single-view human performance capture with cloth simulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, jun 2019. 2

[60] Jinlong Yang, Jean-Sébastien Franco, Franck Hétroy-Wheeler, and Stefanie Wuhrer. Estimation of human body shape in motion with wide clothing. In *European Conference on Computer Vision*, 2016. 3

[61] Jinlong Yang, Jean-Sébastien Franco, Franck Hétroy-Wheeler, and Stefanie Wuhrer. Analyzing clothing layer deformation statistics of 3d human motions. In *European Conf. on Computer Vision*, pages 237–253, 2018. 3

[62] Mihai Zanfir, Alin-Ionut Popa, Andrei Zanfir, and Cristian Sminchisescu. Human appearance transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5391–5399, 2018. 2

[63] Ming Zeng, Jiaxiang Zheng, Xuan Cheng, and Xinguo Liu. Templateless quasi-rigid shape modeling with implicit loop-closure. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 145–152, 2013. 2

[64] Chao Zhang, Sergi Pujades, Michael Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3D scan sequences. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3, 5

[65] Bin Zhou, Xiaowu Chen, Qiang Fu, Kan Guo, and Ping Tan. Garment modeling from a single image. In *Computer graphics forum*, volume 32, pages 85–91. Wiley Online Library, 2013. 2

[66] Qian-Yi Zhou and Vladlen Koltun. Color map optimization for 3d reconstruction with consumer depth cameras. *ACM Transactions on Graphics*, 33(4):155, 2014. 2

[67] Shizhe Zhou, Hongbo Fu, Ligang Liu, Daniel Cohen-Or, and Xiaoguang Han. Parametric reshaping of human bodies in images. In *ACM Transactions on Graphics*, volume 29, page 126. ACM, 2010. 2

[68] Silvia Zuffi and Michael J Black. The stitched puppet: A graphical model of 3d human shape and pose. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 3537–3546. IEEE, 2015. 2